CLAIMS:

1. An automated identification methodology for identification of table of content links in a document comprising:

searching page data to create a list of links in the document;

analyzing each link in conjunction with each other link in the list of links to identify link pairings;

assembling link pairings in order to form clusters of links; and,

examining the links in the cluster of links for locality.

2. The method of claim 1 wherein the step for analyzing each link further comprises determining a score for each link pairing.

3. The method of claim 2 wherein the scoring is determined by a proximity criteria.

4. The method of claim 2 wherein the scoring is determined by a similarity criteria.

5. The method of claim 2 wherein the scoring is determined by a regularity criteria.

6. A system identification methodology for assembling a hyperlinked document comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

analyzing each link in conjunction with each other link to identify link pairings;

assembling link pairings in order to form clusters of links; and,

examining the links in the cluster of links for locality;

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled; and,

performing a document-level analysis that examines the collective set of identified candidate document pages for grouping into one or more documents.

7. The method of claim 6 wherein the step for analyzing each link further comprises determining a score for each link pairing.

8. The method of claim 7 wherein the scoring is determined by a proximity criteria.

9. The method of claim 7 wherein the scoring is determined by a similarity criteria.

10. The method of claim 7 wherein the scoring is determined by a regularity criteria.

11. A system identification methodology for assembling a hyperlinked document comprising:

performing a page-level link analysis that identifies those hyperlinks on a page linking to a candidate document page further comprising a methodology of:

searching page data to create a list of links in the document;

analyzing each link in conjunction with each other link in the list of links to identify link pairings;

assembling link pairings in order to form clusters of links; and,

examining the links in the cluster of links for locality

performing a recursive application of the page-level link analysis to the linked candidate document page and any further nested candidate document pages thereby identified, until a collective set of identified candidate document pages is assembled; and,

performing a document-level analysis that examines the collective set of identified candidate document pages for grouping into one or more documents.

12. The method of claim 11 wherein the step for analyzing each link further comprises determining a score for each link pairing.

13. The method of claim 12 wherein the scoring is determined by a proximity criteria.

14. The method of claim 12 wherein the scoring is determined by a similarity criteria.

15. The method of claim 12 wherein the scoring is determined by a regularity criteria.